Mining Social Media Network using Big Data Analytics to Support Decisions in Educational Organization

[1]Preet Navdeep, [2]Dr. Neeraj Sharma, [3]Dr.Manish Arora
[1]Research Scholar, [2]Professor & Dean, [3]Additional Director (NIELIT)
[1]Department of Computer Applications
[1]IKGPTU, Jalandhar, India

Abstract: Educational data mining and learning analytics are used to research and build models in several areas that can influence Education Organizations. Higher education organizations are beginning to use analytics for improving the services they provide and for increasing student grades and retention. This paper explores different data mining approaches using Big data Analytics techniques which can be applied on Educational data to build up a new environment that give new predictions on the data. This study also looks into various phases of Data mining, how data is fetched from social networking sites and various tools that can be used in the study. A brief literature review has also been done on Education Data Mining and a process has been proposed to analyze Social media data.

Keywords: **Data Mining, Big Data Analytics, Education Data Mining, Learning systems;**

1. INTRODUCTION

Big Data Analytics refers to the use of advanced analytic techniques applied on very large and different data sets that include structured, semi-structured, unstructured data and of different sizes. Big data is the term applied to data sets whose size and type is beyond the ability of traditional relational databases to capture, manage and process the data. Big data is broadly defined as "An accumulation of data that is too large and complex for processing by traditional database management tools" (A.G. Picciano,2012). Big data is defined in terms of 3 V's i.e. Volume, Velocity and Variety [Gartner]. This big data are analyzed using different Analytical Techniques referred to as Big Data Analytics. Big data analytics is used to examine huge amounts of data and can be used to decrypt cipher texts, correlating previously not known variables, finding the trends in the market, checking preferences of customers and finding out data about various businesses and institutions. Data professionals perform analytic operations on the large amount of data which is unconquerable by conventional operations and methodology. Using advanced analytics techniques, government or business agencies can analyze previously untapped data sources independent or together with their existing enterprise data to gain new insights resulting in significantly better and faster decisions.

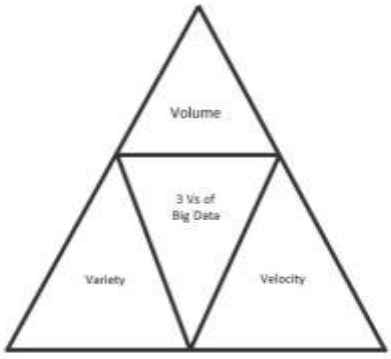

Figure1. 3 V's of Big Data [Gartner]

1.1 Big Data Analytics in education

Technology has been introduced into higher education to improve different practices like teaching, learning. When Big Data is associated with educational objectives and standards, the impact is reflective. Sentiment Data Analysis (Pang, B., & Lee, L. ,2008) is one of the prominent and widely used domains by the research scholars and practitioners. There are number of tools and technologies available for fetching the live datasets, tweets, emotional attributes. Using these tools, the real time tweets and messages can be extracted from Twitter, facebook, whatsapp and many other social media portals. Following steps are carried out to fetch the data from social media sites and analyze the fetched data-

i. Firstly, live data feeds are fetched from social media
ii. Then, Data Cleaning or Refinement of data is done.
iii. Next, Identification of Feature Points is performed.
iv. Then, Mandatory Aspects of data are extracted.
v. Then, Algorithm for Predictive Analytics is implemented.
vi. Next, Investigation of Tweets for classification and predictions is done.
vii. Then, Analysis of Popularity Score or simply Opinion Mining is performed
viii. Last, Detailed Analytic Report is prepared.

By this approach, the emotional attributes of the Internet users on social media portals can be analyzed and prediction can be done. Extracting education data using various data mining techniques is referred to as **Education Data Mining**.

1.1 EDUCATIONAL DATA MINING

Educational Data Mining (EDM) (Baker, R. S., & Yacef, K. ,2009) describes a research field concerned with the application of data mining, machine learning and statistics to information generated from educational institutes (e.g., Universities and Intelligent Tutoring systems). EDM refers to techniques, tools, and research designed for automatically extracting meaning from large repositories of data generated by or related to people's learning activities in educational settings.

1.1.2 Goals of EDM-

Baker and Yacef identified the following four goals of EDM:

1. **Predicting Students' Future Learning Behavior** –This goal can be achieved by creating student models that incorporate the learner's characteristics, including detailed information such as their knowledge, behaviors and motivation to learn. The user experience of the learner and their overall satisfaction with learning are also measured.
2. **Discovering or improving domain models** –Discovery of new and improvements to existing models are possible through various methods and applications of EDM, Examples include illustrating the educational content to engage learners and determining optimal instructional sequences to support the student's learning style.
3. **Studying the effects of educational support** –The effects of educational support can be achieved through learning systems.
4. **Advancing scientific knowledge about learning and learners** –Scientific knowledge about EDM research and the technology can be acquired by building and incorporating student models.

1.1.3 PHASES OF EDUCATIONAL DATA MINING

EDM consists of four phases (Gobert, J. D ,2012):

1. The first phase of the EDM process (not counting pre-processing) is **discovering relationships in data**. This involves searching through a repository of data from an educational environment with the goal of finding consistent relationships between variables.
2. Discovered relationships must then be **validated** in order to avoid over fitting.

3. Validated relationships are applied to make **predictions about future** events in the learning environment.
4. Predictions are used to **support decision-making** processes and policy decisions.

2. REVIEW OF LITERATURE

Bifet, A. et. al (2010) underlines and discusses the challenges that Twitter data streams pose, focusing on classification problems, and then consider these streams for opinion mining and sentiment analysis. To deal with streaming unbalanced classes, we propose a sliding window Kappa statistic for evaluation in time-changing data streams. Using this statistic, the authors perform a study on Twitter data using learning algorithms for data streams.

Barbosa, L. et. al. (2010) propose an approach to automatically detect sentiments on Twitter messages (tweets) that explores some characteristics of how tweets are written and meta-information of the words that compose these messages. Moreover, the work leverage sources of noisy labels as training data. These noisy labels were provided by a few sentiment detection websites over twitter data.

Chen, X. et. al. (2014) developed a workflow to integrate both qualitative analysis and large-scale data mining techniques for analyzing the students' performance and prediction models. The work focused on engineering students' Twitter posts to understand issues and problems in their educational experiences. This work first conducted a qualitative analysis on samples taken from about 25,000 tweets related to engineering students' college life. This work found engineering students encounter problems such as heavy study load, lack of social engagement, and sleep deprivation. Based on these results, a multi-label classification algorithm to classify tweets reflecting students' problems was implemented.

Ana C.E.S Lima and Leandro N de (2012) proposed three approaches for the automatic classification of sentiments, an emotion-based approach, and a word-based approach and a hybrid approach. In the emotion-based approach they used sentiment incorporated in the emotions as criteria to automatically classify the messages. The criteria to select a tweet are the presence of at least one Emoticon. The sentiment is inferred based on the Emoticon. The word-based approach uses words that express sentiment as criteria. In tweets, the presence of words such as good, bad, excellent etc will express sentiment and hence can be inferred. In the hybrid approach a combination of Emoticons and words were used to infer the sentiment. They used Naïve-bayes Algorithm for classifying tweets and concluded that the combined i.e. the hybrid approach yields better results, also, they suggested to add a label "neutral" in future classification

Luo et. al. (2013) highlighted the challenges and an efficient technique to mine opinions from Twitter tweets. Spam and wildly varying language make opinion retrieval within Twitter challenging task.

Luiz F.S Colrtta, nadia F. F. da (2012) consider the problem of standalone Support Vector Machine (SVM) does not give accurate result for finding solution using tweets for a student's experience. So authors suggest that Used combining classifier and cluster ensembles (C3E) to find out students problem so accuracy is improved to find out experience.

Neha R. Kasture, Poonam B. Bhilare(2015)consider the problem is the expression of the verbal throught differs individually, To identifying the right sentiment from the bulk of data becomes the real challenge. Authors suggest that use logical approach to analyze the sentiment of the text available on social media.

Xia et al. (2011) used an ensemble framework for Sentiment Classification which is obtained by combining various feature sets and classification techniques. In their work, they used two types of feature sets (Part-of-speech information and Word-relations) and three base classifiers (Naive Bayes, Maximum Entropy and Support Vector Machines). They applied ensemble approaches like fixed combination, weighted combination and Meta-classifier combination for sentiment classification and obtained better accuracy.

Balkrishna gokulkrishnan, pavalanathan, Nadarajah (2012) find out the problem of student's learning experience's informal post on twitter. So, authors suggest that apply pre-processing and then chain two or more classifiers to find out positive, negative and neutral tweets of the student. Naïve bayes gives accurate result.

Shargabi and Nusari (2010), proposed Data Mining Techniques to discover vital patterns and calculate contribution of academic performance to help decision making. He has used Clustering (by K-means algorithm), association rules (by Apriori algorithm) and decision trees by (J48 and Id3 algorithms) techniques to build the data model.

3. TOOLS

Various tools used can be-

1. **Data Cleansing tools**- The tools used for cleaning and storing textual data are Google Refine and DataWrangler.

2. **Text analysis tools**—These tools are individual or libraries of tools for analyzing social media data once it has been scraped and cleaned. These are mainly natural language processing, analysis and classification tools-

- Transformation tools- These are simple tools that can transform textual input data into tables, maps, charts (line, pie, scatter, bar, etc.), timeline or even motion (animation over timeline), such as Google Fusion Tables, Zoho Reports, Tableau Public or IBM's Many Eyes.
- Analysis tools-These are more advanced analytics tools for analyzing social data, identifying connections and building networks, such as Gephi (open source) or the Excel plug-in NodeXL.

3. **Data Analytics tools**-

- **Cassandra/Hive**-Cassandra is an open source (noSQL) distributed DBMS providing structured 'key-value' store. Key-value stores allow an application to store its data in a schema-less order.
- **Hadoop**- Hadoop is a Java based programming framework that supports the processing of large data sets in a distributed computing environment. An application is broken into different small parts called fragments or blocks that can be run on systems with thousands of nodes.

4. PROPOSED PROCESS OF ANALYZING SOCIAL MEDIA DATA

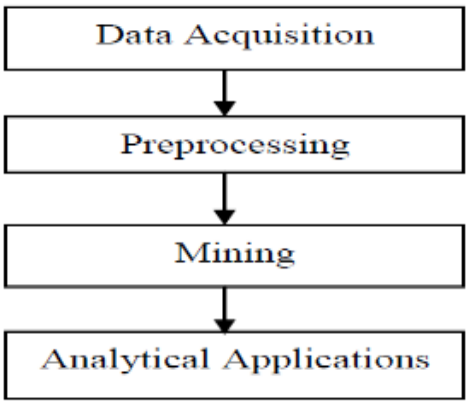

Figure 2. Process of analyzing text

i) **Data Acquisition:** In this data acquisition, data are gathered from different relevant sources such as twitter tweets, online review, newsfeeds and other social media sites etc.

ii) Preprocessing: It is used to remove noisy, inconsistent and incomplete data. For doing the classification, text preprocessing and feature extraction is a primary phase. It has three steps:

- Tokenization
- Removal of stop words
- Stemming- It is a process to reduce derived words to their origin word stem.

iii) Data Mining: Applying different mining techniques to derive usefulness about stored information.

iv) Analytical Application: It provides valuable things from text mining so that it can provide information that helps in improving decision and processes.

V.CONCLUSION

Big Data Analytics in higher education can lead to a transformation in administration, teaching and learning process. With the use of various Big Data Analytics tools, problems of students can be understood and analyzed to make better decisions. There are increasing research interests in using data mining in education. Social networking sites consist of large amount of data related to students' thoughts, views and feedback that can be easily fetched using various APIs. This new emerging domain, called Educational Data Mining, concerns with developing methods that discover knowledge from data originated from educational environments. Data mining is a tremendously vast area that includes employing different techniques and algorithms for pattern finding.

REFERENCES

[1] A.G. Picciano (2012)," The Evolution of Big Data and Learning Analytics in American Higher Education", Journal of Asynchronous Learning Networks, pp 9-20.

[2] e-Governance transactions in 2013 increased manifolds compared to 2012", http://www.thinkalytic.com/2014/e-Governance-transactions-in-2013-increased manifolds-compared-to -2012/, Jan 2, 2014.

[3] WullianallurRaghupathi & VijuRaghupathi," Big Data Analytics Architectures, frameworks, and Tools"

Pang, B., & Lee, L. (2008)," Opinion mining and sentiment analysis. Foundations and Trends® in Information Retrieval", 2(1–2), 1-135.

[4] Baker, R. S., &Yacef, K. (2009)," The state of educational data mining in 2009: A review and future visions", JEDM-Journal of Educational Data Mining, 1(1), 3-17.

[5] Gobert, J. D., Sao Pedro, M. A., Baker, R. S., Toto, E., &Montalvo, O. (2012).,"Leveraging educational data mining for real-time performance assessment of scientific inquiry skills within microworlds", JEDM-Journal of Educational Data Mining, 4(1), 111-143.

[6] Bifet, A., & Frank, E. (2010, October)" Sentiment knowledge discovery in twitter streaming data" In International Conference on Discovery Science (pp. 1-15). Springer Berlin Heidelberg.

[7] Barbosa, L., & Feng, J. (2010, August)" Robust sentiment detection on twitter from biased and noisy data" In Proceedings of the 23rd International Conference on Computational Linguistics: Posters (pp. 36-44). Association for Computational Linguistics.

[8] Chen, X., Vorvoreanu, M., & Madhavan, K. (2014)" Mining social media data for understanding students' learning experiences" IEEE Transactions on Learning Technologies, 7(3), 246-259.

[9] Lima, A. C., and Leandro N. d. C., "Automatic sentiment analysis of Twitter messages." Computational Aspects of Social Networks (CASoN),2012 Fourth International Conference on. IEEE, 2012.

[10] ZhunchenLuo, Miles Osborne, TingWang, An effective approachto tweets opinion retrieval", Springer Journal onWorldWideWeb,Dec 2013.

[11] Luiz F.S Colrtta,nadia F. F. da silva,"Combining classification and cluster for tweet sentiment analysis." 2014 IEEE.

[12] Neha R. Kasture, Poonam B. Bhilare, An Approach for sentiment analysis on social networking sites, 2015 International Conference on Computing Communication Control and Automation, 2015 IEEE.

[13] Xia, R., Zong, C., and Li, S., "Ensemble of feature sets and classification algorithms for sentiment classification," Information Sciences: an International Journal, Vol. 181, Issue 6, pp. 1138–1152, 2011.

[14] Balkrishnagokulkrishnan, pavalanathan,nadarajah prasath, "Opinion mining and sentiment analysis on a twitter data stream "2012 IEEE.

[15] Al-shargabi, A. & Nusari, A. Discovering vital patterns from UST students data by applying data mining techniques Computer and Automation Engineering (ICCAE), 2010. The 2nd International Conference on, 2010, 2, 547-551.